Vol.2017-HCI-175 No.3
Vol.2017-UBI-56 No.3
2017/11/1

スピーチ支援のための広頚筋隆起センシングに基づく良い開口での母音発話認識

西村幸泰[†1]　橋田朋子[†1]

概要：スピーチは言語的伝達能力と非言語的伝達能力から評価されており，特に開口はこの両方を補助するという重要な役割を持つ．開口は母音発話の際に主に行われる．開口の中でも，話者の明瞭な発音を促し表情を豊か見せることができるような開口を本研究では良い開口と定義する．良い開口であるかどうかは首にある広頚筋という筋肉の隆起から推測できると筆者らは考えた．そこで本研究では，良い開口での母音発話の支援を目指し，良い開口での母音発話を認識するシステムを提案する．具体的には，広頚筋の隆起をフォトリフレクタで測定し，そのデータを機械学習にかけることによって，良い開口で母音発話が行われているかどうかを判断するシステムを実装する．提案システムの精度計測実験を行い，その結果を報告する．最後に良い開口での母音の発話のフィードバックを返すアプリケーションについて述べる．

キーワード：パブリックスピーチ，スピーチ支援，開口，機械学習，

Recognition of Vowel Utterance Produced by Good Mouse Shape by Sensing Platysma Muscle Building for Public Speaking

YUKIHIRO NISHIMURA[†1]　TOMOKO HAHSIDA[†1]

Abstract: Speech is evaluated by verbal delivery and non-verbal delivery．Mouse shape supports both deliveries，which makes itself important．Mouse shape changes mainly when people product vowel utterance．Good mouse shape helps making clear pronounce and clear facial expression．We assume good mouse shape can be estimated by platysma muscle building，which exists in a neck．We propose a system that recognizes vowel produced by good mouse shape．We implement this system by sensing platysma muscle building with photoreflector and by analyzing sensor data with machine learning．We report the result of accuracy measurement experiment of the proposed system and report the result．Finally，we describe the application which returns feedback of vowel utterance with good opening.

Keywords: Public Speech，Presentation Training，Mouse Shape，Machine Learning

1. はじめに

シャドウイング練習の支援[1]や人前でのスピーチ練習の支援[2]などスピーチ支援に関する研究は盛んに行われている．スピーチ支援の研究の中でもスピーチの評価に関する研究領域がある．パブリックスピーチの分野ではスピーチは言語的伝達能力と非言語的伝達能力の 2 点から評価されている[3][4]. 言語的伝達能力は，発音の明瞭さや抑揚，声の高さ，話す速さなどの音声表現の多様性に基づき評価される．言語的伝達能力に関する研究として例えば，あらかじめ録音された話者の音声の抑揚や声量を操作して再生するものや[5]，スピーチ中の抑揚や声量や間の取り方から話者の感情を認識する研究[6]がある．一方，非言語的伝達能力は，身振りやアイコンタクト，表情などの音声以外の表現の豊かさに基づき評価される．非言語的伝達能力に関する研究として例えば，録画した話者の映像から身振りや表情の豊かさをフィードバックするもの[7]や，聴衆の注意を散漫させるような話者の無意識な身体の揺れや手の動きを警告するもの[8]などがある．これまでのスピーチ支援に関する研究では言語的伝達と非言語的伝達のどちらか一方のみが支援の対象とされてきた. 本論文では 話者の開口に着目し，2 種類の伝達を共に支援することを目指す.

開口は主に母音を発話するときに行われる．母音の種類によって，明瞭に発話するために適した口の開き具合は異なっている[9]. 例えば/u/の音の発話では唇を強くすぼめるのが良いとされており，その他の母音の発話では横方向や下方向に大きく唇を開くのが良いとされている．本研究では，これらの明瞭な母音を発話できる口の開き具合を「良い開口」と定義する．話者が良い開口であるとき，聴衆は口の動きの変化を大きく感じるため，話者の表情が豊かであるという印象を得やすくなる．また聴衆は話者の口の動きがはっきりと見えるため，上手く聞き取れなかった場合であっても話者が何と発言したかを推測しやすくなる．そのため，良い開口は発音を明瞭にするという点で言語的伝

†1 早稲田大学
Waseda University.

Vol.2017-HCI-175 No.3
Vol.2017-UBI-56 No.3
2017/11/1

図 1　広頚筋の位置

達能力を補助するだけでなく，表情を豊かにするという点で非言語的伝達能力をも補助するため，スピーチにおける開口は非常に重要であるといえる．

しかしスピーチを支援するこれまでの研究領域では，話者の開口具合はあまり注目されてこなかった．筆者は学士時代に英語スピーチのサークルに所属しており，全国大会の決勝に数回出場する中で，スピーチにおける開口の重要性に気が付き，良い開口を支援したいと考えた．ここで，筆者の主観では，良い開口ができた際は首の筋肉が上手く使用できている．首には広頚筋という筋肉が存在する．広頚筋は口角を横方向に動かすときと顎を下に引くときに隆起するという特徴を持つ．この性質から良い開口のときは広頚筋が隆起すると推測した．

以上より，スピーチの上手さは言語的伝達能力と非言的伝達能力から規定され，良い開口はそれらの両方に良い影響を与える．開口は主に母音発話の際に行われ，良い開口での母音発話と首の筋肉の隆起が関係を持つと示唆される．そこで筆者らは首の筋肉の隆起を測定することにより，良い開口での母音発話が認識できるのではないかと考えた．具体的には表情筋を動かす首の筋肉の隆起をフォトリフレクタで測定し，そのデータを機械学習にかけることによって，良い開口での母音の発話を判断する仕組みを提案する．本稿では提案システムの詳細と精度計測実験の結果，及び良い開口での母音の発話のフィードバックを返すアプリケーションについて述べる．

2. 関連研究

話者の発話を認識する研究領域に注目すると大きく分けて 2 つの領域がある．一つは外部機器による非接触型センシングを用いる領域で，もう一方はウェアラブルデバイスによる接触型センシングを用いる領域である．非接触型センシングを用いる研究の例としては，口唇動作を外部カメラによって測定することで発話を認識する LipNet[10]や，マイクによって話者の音声を測定することで発話を認識する Lopez らの研究[11]が挙げられる．これらの研究は外部機器によるセンシングのため，明るさや騒音といった環境の変化に対して弱いといった問題がある．対して接触型センシングを用いる研究の例としては，声帯付近の首表面の振動を加速度センサによって測定することで発話を認識する Daryush らの研究[12]や，胸壁表面の皮膚振動を振動センサによって測定することで発話を認識する Sundberg らの研究[13]がある．これらの研究はウェアラブルデバイスを用いているため，環境の変化に対して強いという利点がある．しかしスピーチならではの顔の表情といった非言語的伝達の情報を排除しているといった問題がある．またセンサ部分が大きいため装着時のユーザへの違和感が大きいという問題がある．

そこで本研究ではウェアラブルデバイスを用いて，言語的伝達と非言語的伝達の両方の支援を目指し，1 種類のセンサから言語的伝達と非言語的伝達の情報を取得し，話者の良い開口での母音発話を認識するシステムの実装をする．また，センサとして，小型で実装がコンパクトになりやすいフォトリフレクタを選択する．

3. システム

3.1 システム概要

広頚筋を接触型センシングで測定し，良い開口での母音発話を認識するシステムを提案する．バンド状に配置したフォトリフレクタモジュールを首に装着し，広頸筋の隆起を測定することで発声したどのような開口で母音発話をしたかを認識するシステムを実装した．実装にあたりスピーチ時に話者への装着の違和感を減らし，開口の動作を阻害しないために，デバイスの装着場所として口周りでなく首を選択した．システム構成を図 2 に示す．

3.2 ハードウェア

本システムは 8 個のフォトリフレクタモジュール，マイコン（Arduino Mega），PC から構成される．制作したフォ

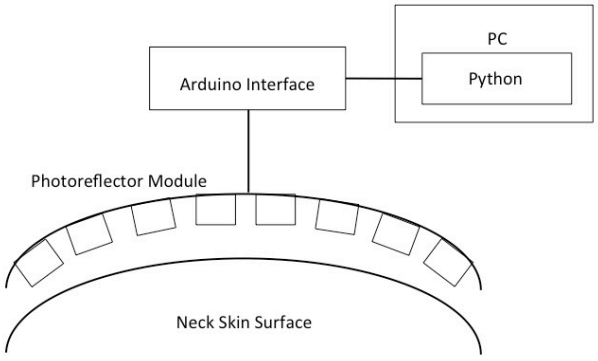

図 2 システム構成

Vol.2017-HCI-175 No.3
Vol.2017-UBI-56 No.3
2017/11/1

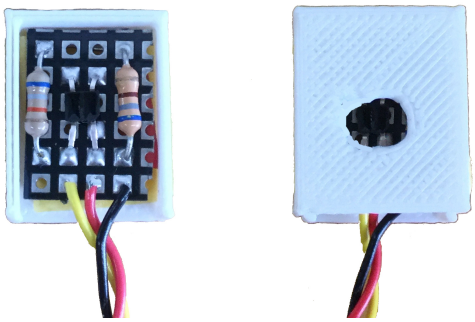

図 3 フォトリフレクタモジュールの内観と外観

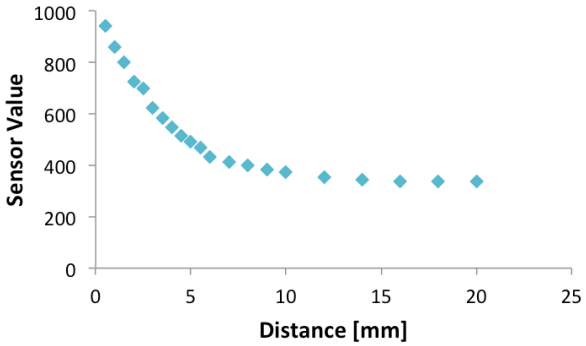

図 4 モジュールのセンサ特性

トリフレクタモジュールを図 3 に示す．フォトリフレクタモジュールは杉浦らの研究[14]を参考にした．モジュールの寸法は 1.7cm×2.0cm，厚さ 7.5mm である．フォトリフレクタは GENIXTEK CORP.の TPR-105F を使用した．モジュールの位置の固定のためにネックバンドを使用した．モジュール間の距離は 20mm とした．8 個のモジュールの平均をとったセンサ特性を図 4 に示す．それぞれの計測においては 10 サンプルを収集しその平均をとった．グラフの縦軸はセンサモジュールの入力値，横軸はセンサモジュールから皮膚への距離である．センサモジュールの入力値とセンサモジュールから皮膚への距離の対応関係は非線形であり，近距離での計測が可能であることを確認した．このセンサは近距離であるほどより高い分解能をもつことが確認された．

3.3 ソフトウェア

Arduino 環境下で，各フォトリフレクタモジュールからの入力データをシリアル通信で PC に送信した．PC では送られてきたデータを Python のプログラムを用いて機械学習した．機械学習の学習器として Support Vector Machine（SVM）を用いた．ユーザが本システムを使用する際には，覚えさせたい開口（ニュートラルな表情での無発声状態，良い開口での各母音発話状態）で発話し，その都度センサの入力データを収集する．1 回の学習毎にフレームレート 60 の間隔で 10 個のデータを収集し，この学習を 10 試行繰り返す．合計では，600 個のデータ（6 種類の開口×10 個のデータ ×10 試行）が集められる．収集したデータを用いて，それぞれの開口での発話を認識するための識別モデルを個別に生成する．これらのモデルを用いて行われる認識は実時間で実行することが可能である．

4. 実験

4.1 概要

実装したシステムの精度の評価を目的とした．具体的には，本システムを用いて広頚筋の隆起のセンシングからどのような開口具合で母音が発声されたかを推定し，その認識率を調べた．

4.2 認識セット

今回の実験で認識を行った開口の一覧を図 5 に示す．認識セットはニュートラルな表情時の無発声状態，ニュートラルな表情時の母音発声状態および良い開口での母音発声状態の計 11 種類から構成される．認識セットは [9]を参考に設定した．

4.3 手続き

この実験は 20 代男性 2 人と女性 1 人の計 3 人の被験者で行われた．この実験は室内環境で行われた．実験手順は以下の通りであった．

(1) 被験者はまず椅子に座り，バンド上に配置したフォトリフレクタモジュールを首に装着した．センサの配置した位置を図 6 に示す．喉仏の下から首の付け根付近において広頚筋の隆起を測定した．

(2) 被験者は，実験者の指示に従って認識セットに含まれる全ての開口での母音発話を練習した．練習の際，図 4 の資料を見せながら主に良い開口での発話を指導した．

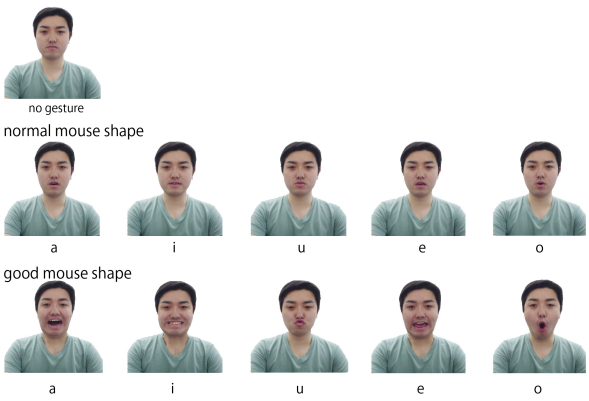

図 5 認識セット一覧

図 6 センサの配置

(3) 被験者はニュートラルな表情時の無発声状態をした後に，実験者の指示した開口での発話をし，実験データの収集をした．それぞれの開口データを記録する際，1 回の試行毎にフレームレート 60 の間隔で 10 個のデータを収集した．合計では，1100 個のデータ(11 種類の開口×10 個のデータ ×10 試行) を全ての被験者から集めた．

4.4 結果

4.4.1 被験者毎の認識率

10 分割交差検証を用いて，各開口での被験者毎の認識率を算出した．具体的には，1 人の被験者から収集した 1 つの動作の実験データについて，その中の 9 試行分のデータを教師データとして学習し，その識別モデルを生成した．残り 1 試行分のデータをテストデータとして，生成したモデルの精度を算出した．全ての教師データとテストデータの組み合わせに対して同様の操作を行い，得られた複数個の認識率の平均値を計算する．この値を，ある 1 人の被験者における本システムの認識率とした．3 人の被験者それぞれから得られた認識率の平均値を計算し，本システムの被験者毎の認識率とした．

上記の分析を行った結果，被験者毎の認識率の平均は全体で 83.3%（SD = 10.4）であり，ニュートラルな表情時の無発声状態で 100%，ニュートラルな表情で 76.0%（SD = 9.9），良い開口で 87.3%（SD = 7.6）であった．開口毎における被験者毎の認識率を表 1 に示す．

4.4.2 被験者間の認識率

本システムの一般性を調べるため，3 人の被験者から収集した実験データを全て合わせて学習とテストを行い，被験者間の認識率を算出した．1 つの動作の実験データについて，2 人分の被験者の実験データを全て教師データとして学習し，その動作の識別モデルを生成する．残り 1 人の被験者の実験データを全てテストデータとして，生成したモデルの精度を算出した．全ての教師データとテストデータの組み合わせに対して同様の操作を行った．得られた複数の精度の平均値を計算し，対象の動作の被験者間の認識率とした．

表 1 被験者毎の認識率

		Neurtral Mouse Shape						Good Mouse Shape					
	Actual / Predict	normal	a[%]	i[%]	u[%]	e[%]	o[%]	a[%]	i[%]	u[%]	e[%]	o[%]
Neutral Mouse Shape	normal	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	a	0.7	64.7	0.0	10.7	10.0	10.0	4.0	0.0	0.0	0.0	0.0
	i	0.0	0.0	92.7	0.0	0.0	0.0	0.0	7.3	0.0	0.0	0.0
	u	0.0	6.7	0.0	76.0	17.3	0.0	0.0	0.0	0.0	0.0	0.0
	e	3.3	1.7	0.0	9.0	79.3	3.3	3.3	0.0	0.0	0.0	0.0
	o	0.0	15.0	0.0	1.0	5.7	67.7	10.7	0.0	0.0	0.0	0.0
Good Mouse Shape	a	0.0	6.0	0.0	0.0	0.0	6.3	87.7	0.0	0.0	0.0	0.0
	i	0.0	0.0	0.0	0.0	0.0	0.0	0.0	93.3	0.0	0.0	6.7
	u	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	83.7	0.0	13.0
	e	0.0	0.0	0.0	3.3	0.0	0.0	0.0	0.0	0.0	96.7	0.0
	o	0.0	0.0	3.3	0.0	0.0	0.0	0.0	13.3	8.3	0.0	75.0

表 2 被験者間の認識率

		Neurtral Mouse Shape						Good Mouse Shape				
	Actual / Predict	normal	a[%]	i[%]	u[%]	e[%]	o[%]	a[%]	i[%]	u[%]	e[%]	o[%]
Neutral Mouse Shape	normal	46.7	6.7	0.0	5.0	0.0	0.0	0.0	0.0	33.3	8.3	0.0
	a	0.0	6.7	0.0	0.0	1.3	6.7	52.0	0.0	33.3	0.0	0.0
	i	0.0	0.0	66.7	0.0	19.3	0.0	0.0	14.0	0.0	0.0	0.0
	u	0.0	0.0	0.0	29.0	4.3	11.3	26.7	0.0	22.0	0.0	0.0
	e	0.0	0.0	0.0	25.3	27.0	0.0	14.3	0.0	33.3	0.0	0.0
	o	4.7	1.0	0.0	2.7	3.0	22.0	27.7	0.0	39.0	0.0	0.0
Good Mouse Shape	a	0.0	3.7	0.0	2.0	2.3	3.7	55.0	0.0	33.3	0.0	0.0
	i	0.0	0.0	0.0	30.0	0.0	0.0	0.0	52.7	9.0	0.0	5.0
	u	15.3	0.0	0.0	9.7	3.3	3.3	8.3	24.3	12.3	20.3	3.0
	e	0.0	0.0	0.0	13.3	20.3	3.3	0.0	6.7	36.3	3.3	16.7
	o	6.7	0.0	0.0	20.0	0.0	0.0	0.0	26.3	6.7	16.7	23.7

被験者間の認識率を表 2 に示す．被験者間の認識率の平均は全体で 31.4%（SD = 20.1）であり，ニュートラルな表情時の無発声状態で 46.7%，ニュートラルな表情で 30.3%（SD = 19.8），良い開口で 21%（SD = 21.0）であった．

4.5 考察

4.5.1 被験者毎の認識率

ニュートラルな表情時の認識率よりも良い開口時の方が認識率は高かった．また，誤認識に関して，ニュートラルな表情では別のニュートラルな表情の母音に，良い開口での誤認識は別の良い開口での母音にほとんどが判定されていた．このことからニュートラルな表情での開口と良い開口かの判断ができると示唆された．

4.5.2 被験者間の認識率

被験者間の認識率はいずれも低い値に留まり，認識の一般性に課題が残る結果となった．原因は 2 つあると考える．1 つめの原因は特徴量の収集が局所的になったことであり，2 つめの原因は首の形状の違いである．1 つめの原因に関して，今回センサ間隔を 20mm に設定したが，首の断面を正円だと仮定したとき首の中心から約 20deg 間隔でセンサを配置したことになる．この間隔が粗であったため特徴量の収集が局所的になったと考える．精度を上げるためにはより密にセンサを配置し広い範囲での特徴量を収集すると良いと考える．

2 つめは首の形状の違いである．テストデータが女性の場合と男性の場合で大きく異なった．テストデータが男性

のときはものになった際の認識率が著しく低かった．喉仏の有無による影響を考慮し，システムを喉仏の下に付けたとはいえ，無発声時の際のセンサ値が大きく異なっていた．性差を考慮したキャリブレーションの必要性があると判明した．

4.6 実験のまとめ

被験者毎の認識率はニュートラルな表情の母音発話よりも良い開口時での母音発話のほうが高い値となり，90%に近い値となった．これは本システムの有効性を示している．また被験者間の認識率はニュートラルな表情での母音発話および良い開口での母音発話の両方で低い値となった．これは本システムには一般性に課題が残ることを示している．しかしウェアラブルデバイスは基本的に個人で使用することを前提にしているため，大きな問題ではないと考える．実験の結果から本システムは 1 人のユーザに使われる場合に高い認識精度を発揮する．

5. アプリケーション

話者の発話した音素のうち母音部分のみを認識し，良い開口で発話されたかどうかをディスプレイでフィードバックするアプリケーションを制作した．図 7 に実装したアプリケーションを示す．使用方法は次の通りである．このフィードバックは実時間で行われる．

(1) まずユーザは机の前の椅子に座り，図 6 に示した首の位置に本システムを装着する．
(2) 各母音にてニュートラル表情で無発声の状態とニュートラルな表情での開口と良い開口を本システムに学習させ，それらの教師データを作成する．
(3) ユーザは机上に配置されたディスプレイに向かってスピーチの原稿を読む．
(4) ユーザはディスプレイに表示された母音発話認識の結果を見る．認識の結果は良い開口のときにのみ母音発話を判定し，ニュートラルな表情のときは母音を判定しない．例えば良い開口で/ki/と発声した場合，良い開口での/i/と判定される．

6. まとめ

本稿では，首に存在する広頚筋の隆起を測定し機械学習させることによって，良い開口での母音発話の認識を可能にするシステムを提案した．広頚筋の隆起の測定にはフォトリフレクタモジュールを，学習器には SVM を用いた．安静座位かつ首の向きを一定にするという条件の下で実験を行い，本システムはニュートラルな表情時の母音発話よりも良い開口時での母音発話で高い認識率が得られた．実時間で話者の良い開口での母音発話をフィードバックするアプリケーションについて述べた．今後の展望としてスピーチ中どのタイミングで良い開口で発話をしていたかをタイムラインで示すものやスピーチ中の良い開口での発話が占める割合をフィードバックするアプリケーションを制作する．

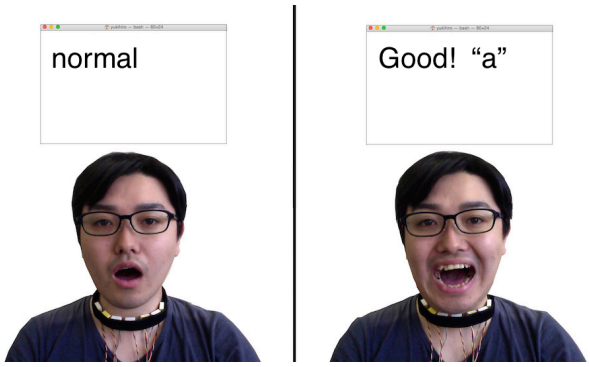

図 7 アプリケーション

参考文献

[1] 張鑫磊，味八木崇，暦本純一: “WithYou: 音声認識を用いたインタラティブシャドーイングコーチ”，インタラクション 2016，2016.
[2] H. Trinh，R. Asadi，D. Edge，T. Bickmore，RoboCOP: A Robotic Coach for Oral Presentations，Proceedings of the ACM on Interactive，Mobile，Wearable and Ubiquitous Technologies，v.1 n.2，p.1-24，June 2017
[3] Swai Johar. Emotion, Affect and Personality in Speech. Springer, p.1-6，2016.
[4] Rodrigues IG，“Verbal and nonverbal signals in face-to-face interaction: a theoretical framework for a holistic micro-analysis (The example of a parenthesis)”. Interacting Bodies，Lyon，15–18 June 2005 .
[5] Michelle Fung，Yina Jin，RuJie Zhao，Mohammed (Ehsan) Hoque，ROC speak: semi-automated personalized feedback on nonverbal behavior from recorded videos，Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing，2015.
[6] M. Iftekhar Tanveer，Ru Zhao，Kezhen Chen，Zoe Tiet，Mohammed Ehsan Hoque，AutoManner: An Automated Interface for Making Public Speakers Aware of Their Mannerisms，Proceedings of the 21st International Conference on Intelligent User Interfaces，2016.
[7] Xiang Li, and Jun Rekimoto, “SmartVoice: A Presentation Support System for Overcoming the Language Barriers”，CHI2014，2014.
[8] Tato R，Santos R，Kompe R，Pardo JM，Space improves emotion recognition. In:Proceedings of international conference on speech language processing (ICSLP).，pp 2029–2032，2002.
[9] “発音練習テキスト”. http://el.minoh.osaka-u.ac.jp/flit/public/en/docs/pronun_ex_text.pdf，(参照 2017-10-10)
[10] Yannis M. Assael，Brendan Shillingford，Shimon Whiteson，Nando de Freitas, “LipNet: End-to-End Sentence-level Lipreading”，arXiv，2016.
[11] Lopez-Cozar R，Araki M，“Spoken，multilingual and multimodal dialogue systems.” Development and assessment. Wiley，West Sussex，2006.
[12] Daryush D. Mehta，Member，IEEE，Jarrad H. Van Stan，and

Vol.2017-HCI-175 No.3
Vol.2017-UBI-56 No.3
2017/11/1

Robert E. Hillman, “Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer”, 2016 IEEE, 2016.

[13] J. Sundberg, “Chest wall vibrations in singers,” J. Speech Hear. Res., vol. 26, no. 3, pp. 329–340, 1983.

[14] 杉浦裕太，稲見昌彦，五十嵐健夫，光透過性を利用した薄い布の伸縮の計測とその応用，バーチャルリアリティ学会論文誌 特集論文 デジタルファブリケーションと VR，Vol.20, No.2，pp.115-121，2015.